Document Classification with Reuter50 News Text: Update 1

Muzhe Zeng
mzeng6@wisc.edu

1 Dataset

The dataset is the subset of RCV1. These corpus has already been used in author identification experiments. In the top 50 authors (with respect to total size of articles) were selected. 50 authors of texts labeled with at least one subtopic of the class CCAT(corporate/industrial) were selected.That way, it is attempted to minimize the topic factor in distinguishing among the texts. The training corpus consists of 2,500 texts (50 per author) and the test corpus includes other 2,500 texts (50 per author) non-overlapping with the training texts. For more details, see this.

2 Current Progress

The python file `dataProcess.py` transform the text dataset into a pandas dataframe and is ready for the machine learning tasks. The link to the Github: `https://github.com/MuzheZeng/Document-Classification-with-vsp`

3 Future Plan

Implement the planned machine learning models to test the performance and compare. Later on, I will fine-tune the models.

2020 Fall, Matrix Methods in Machine Learning (ECE 532) Final Project.